LendingClub

Predictive Models and Decision Trees

Lecture plan

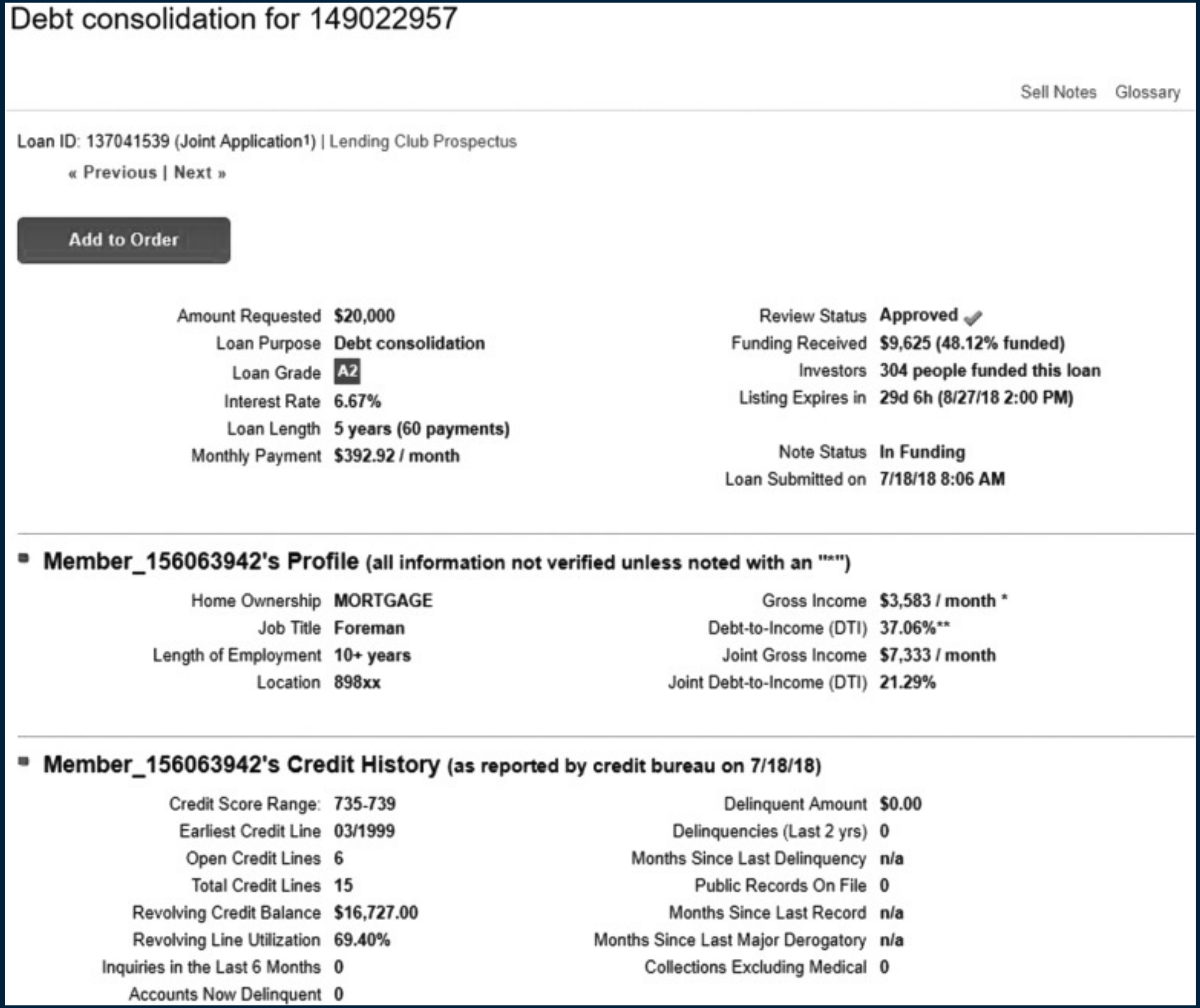
- Predictive models
- Features and targets
- Supervised Segmentation
- Decision Trees
- ID3 algorithm

Predictive Models

- ***Predicti*****v**e models apply statistics to predict outcomes
- The key criteria of evaluating a predictive model is its ability to ***forecast***, typically events in the future

Data available on each loan application

- **We think of these fields of information as attributes or property of the object of interest**



Debt consolidation for 149022957

Loan ID: 137041539 (Joint Application1) | Lending Club Prospectus

« Previous | Next »

Amount Requested	$20,000	Review Status	Approved ✓
Loan Purpose	Debt consolidation	Funding Received	$9,625 (48.12% funded)
Loan Grade	A2	Investors	304 people funded this loan
Interest Rate	6.67%	Listing Expires in	29d 6h (8/27/18 2:00 PM)
Loan Length	5 years (60 payments)		
Monthly Payment	$392.92 / month	Note Status	In Funding
		Loan Submitted on	7/18/18 8:06 AM

Member_156063942's Profile (all information not verified unless noted with an "*")

Home Ownership	MORTGAGE	Gross Income	$3,583 / month *
Job Title	Foreman	Debt-to-Income (DTI)	37.06%**
Length of Employment	10+ years	Joint Gross Income	$7,333 / month
Location	898xx	Joint Debt-to-Income (DTI)	21.29%

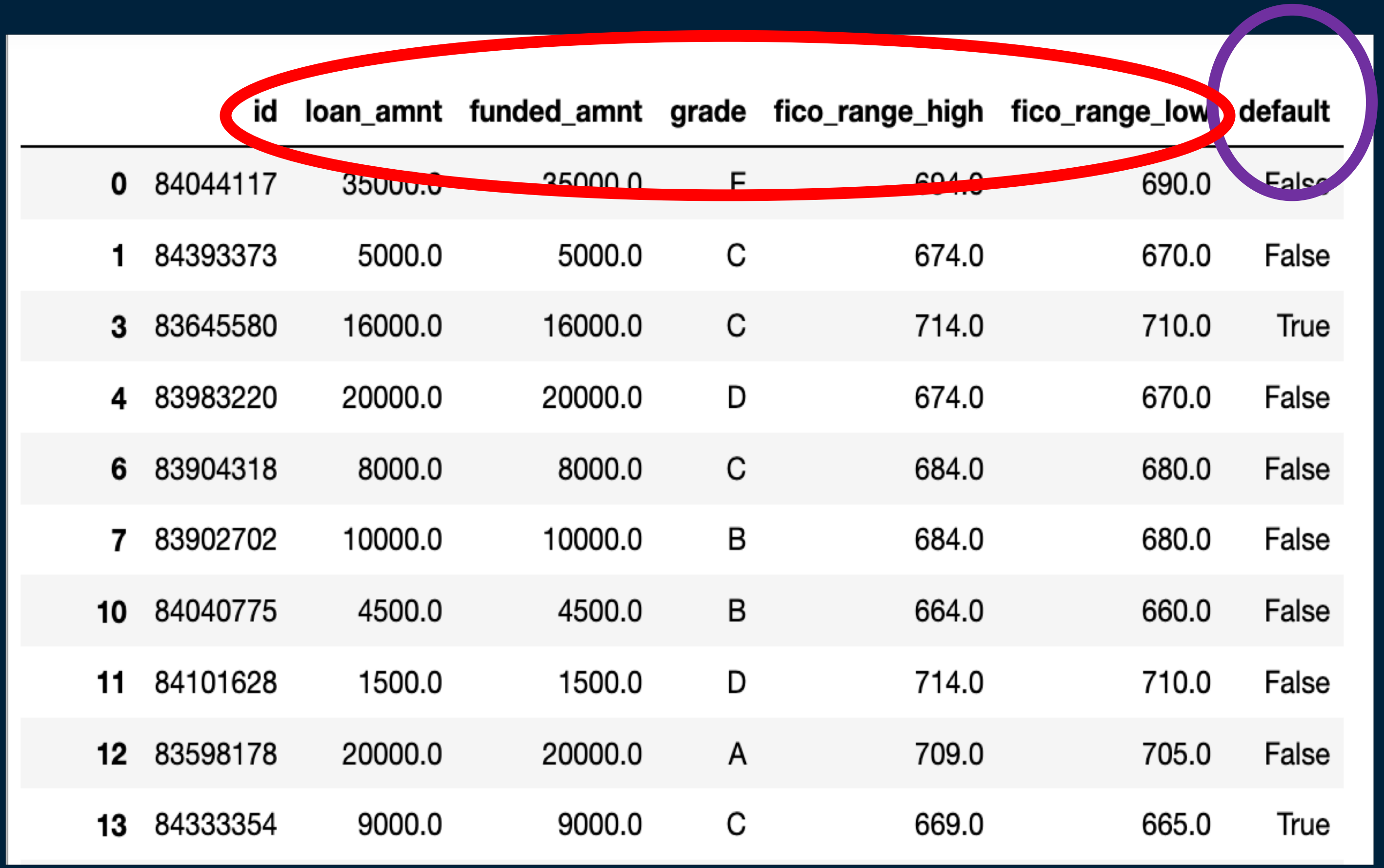
Member_156063942's Credit History (as reported by credit bureau on 7/18/18)

Credit Score Range:	735-739	Delinquent Amount	$0.00
Earliest Credit Line	03/1999	Delinquencies (Last 2 yrs)	0
Open Credit Lines	6	Months Since Last Delinquency	n/a
Total Credit Lines	15	Public Records On File	0
Revolving Credit Balance	$16,727.00	Months Since Last Record	n/a
Revolving Line Utilization	69.40%	Months Since Last Major Derogatory	n/a
Inquiries in the Last 6 Months	0	Collections Excluding Medical	0
Accounts Now Delinquent	0		

Attributes / Features of a model

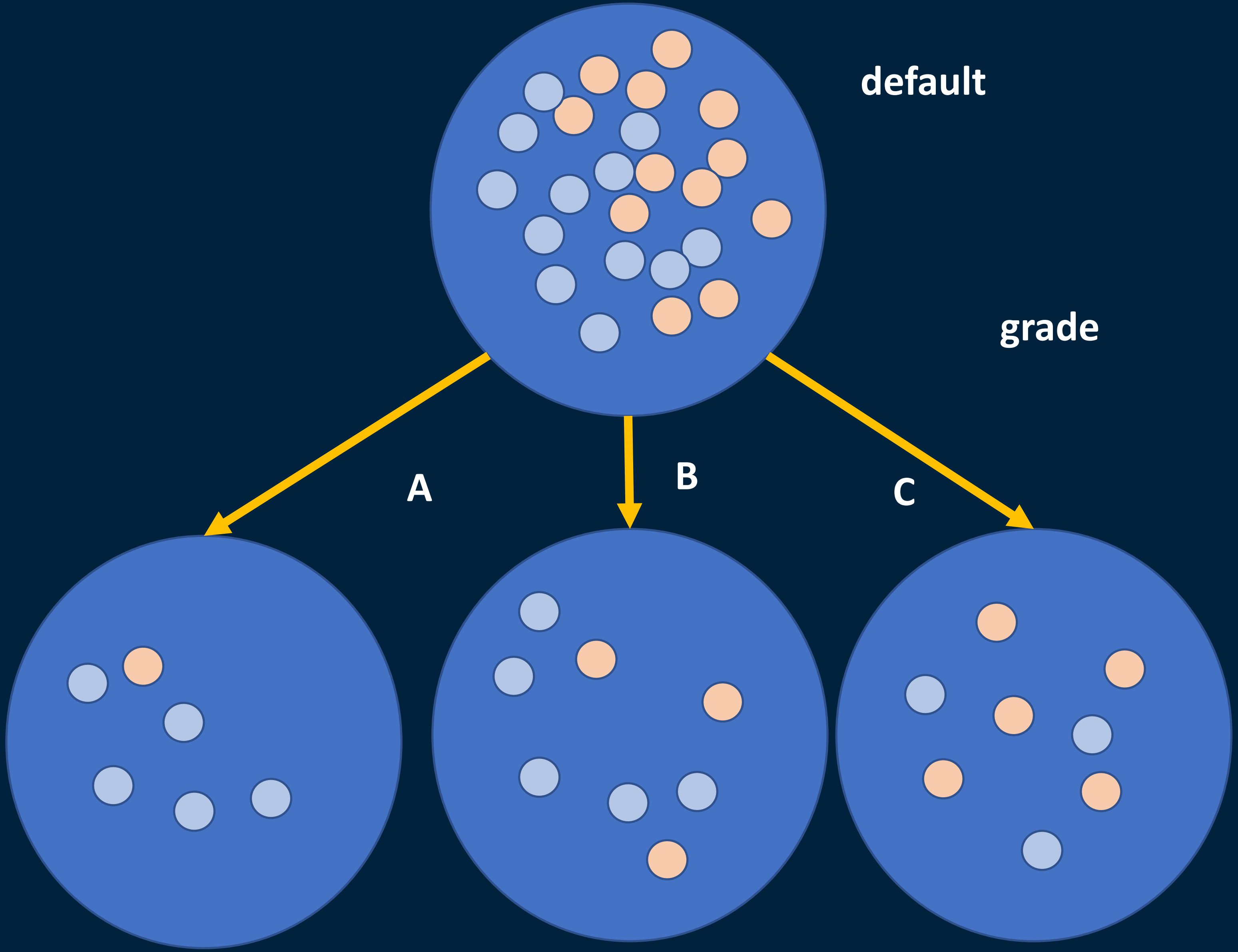
Attribute / independent variables

target attribute / dependent variables



	id	loan_amnt	funded_amnt	grade	fico_range_high	fico_range_low	default
0	84044117	35000.0	35000.0	E	694.0	690.0	False
1	84393373	5000.0	5000.0	C	674.0	670.0	False
3	83645580	16000.0	16000.0	C	714.0	710.0	True
4	83983220	20000.0	20000.0	D	674.0	670.0	False
6	83904318	8000.0	8000.0	C	684.0	680.0	False
7	83902702	10000.0	10000.0	B	684.0	680.0	False
10	84040775	4500.0	4500.0	B	664.0	660.0	False
11	84101628	1500.0	1500.0	D	714.0	710.0	False
12	83598178	20000.0	20000.0	A	709.0	705.0	False
13	84333354	9000.0	9000.0	C	669.0	665.0	True

Informative Attributes



Informative Attributes

Grade

grade	#defaults	default likelihood
A	3041.0	8.47%
B	12448.0	17.97%
C	19888.0	27.98%
D	13154.0	37.89%
E	7479.0	46.04%
F	3450.0	54.43%
G	1061.0	58.91%

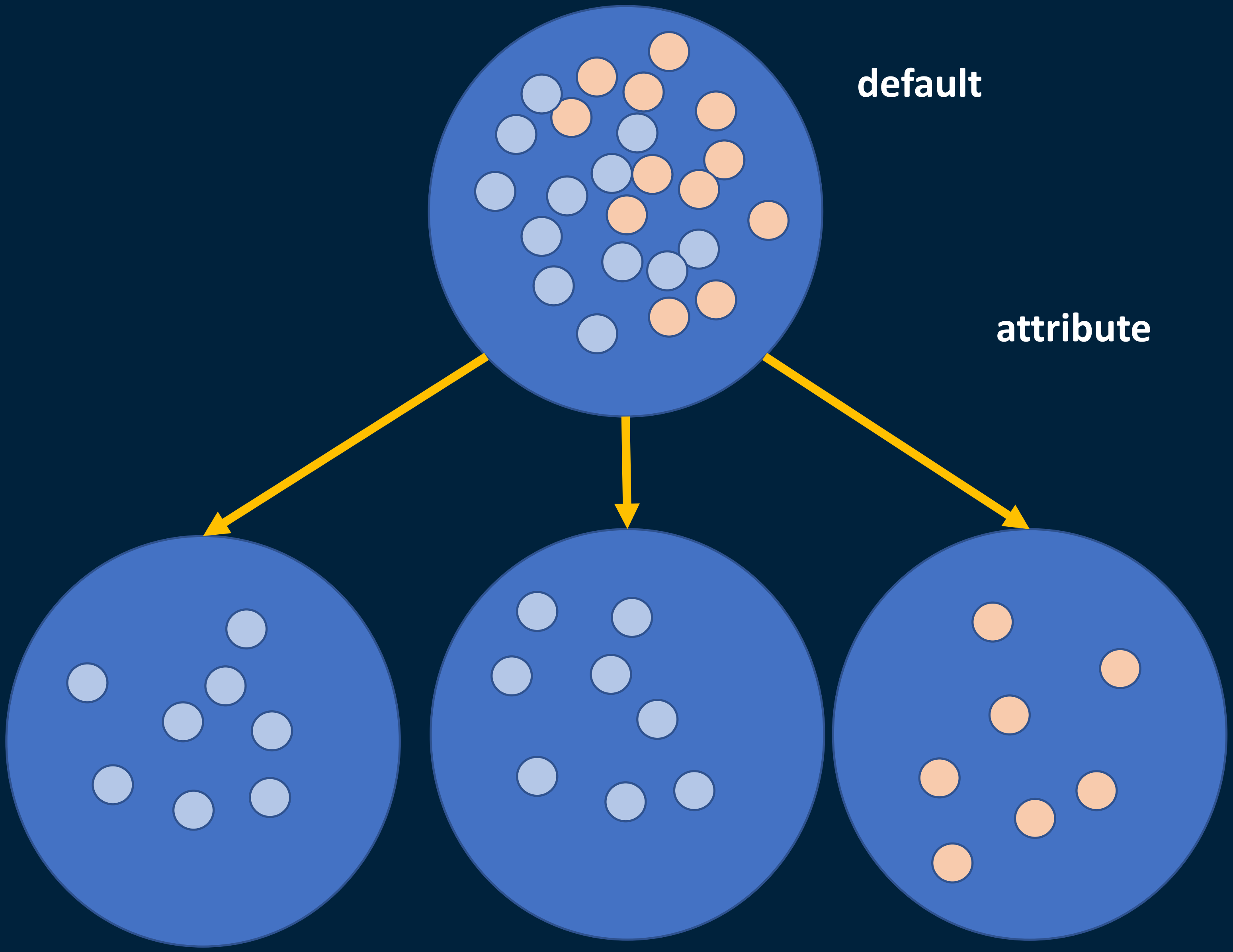
Informative Attributes

Fico

Fico	#defaults	default likelihood
<700	47401.0	29.76
700< <800	12978.0	17.55
>800	142.0	6.66

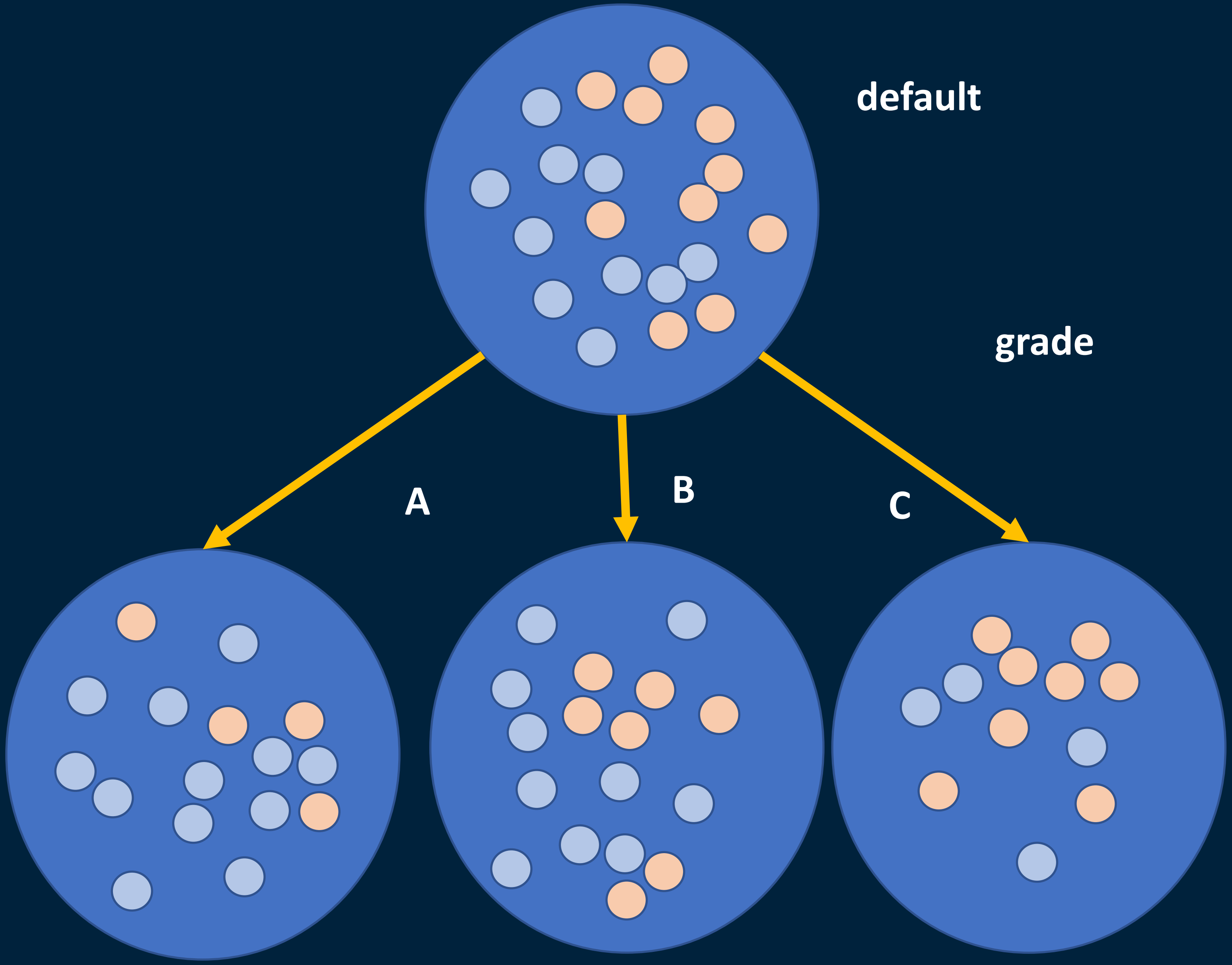
Best case scenario

- The attributes separate perfectly loans that will default and those that won’t
- In other words attributes will separate instances into “pure” sets



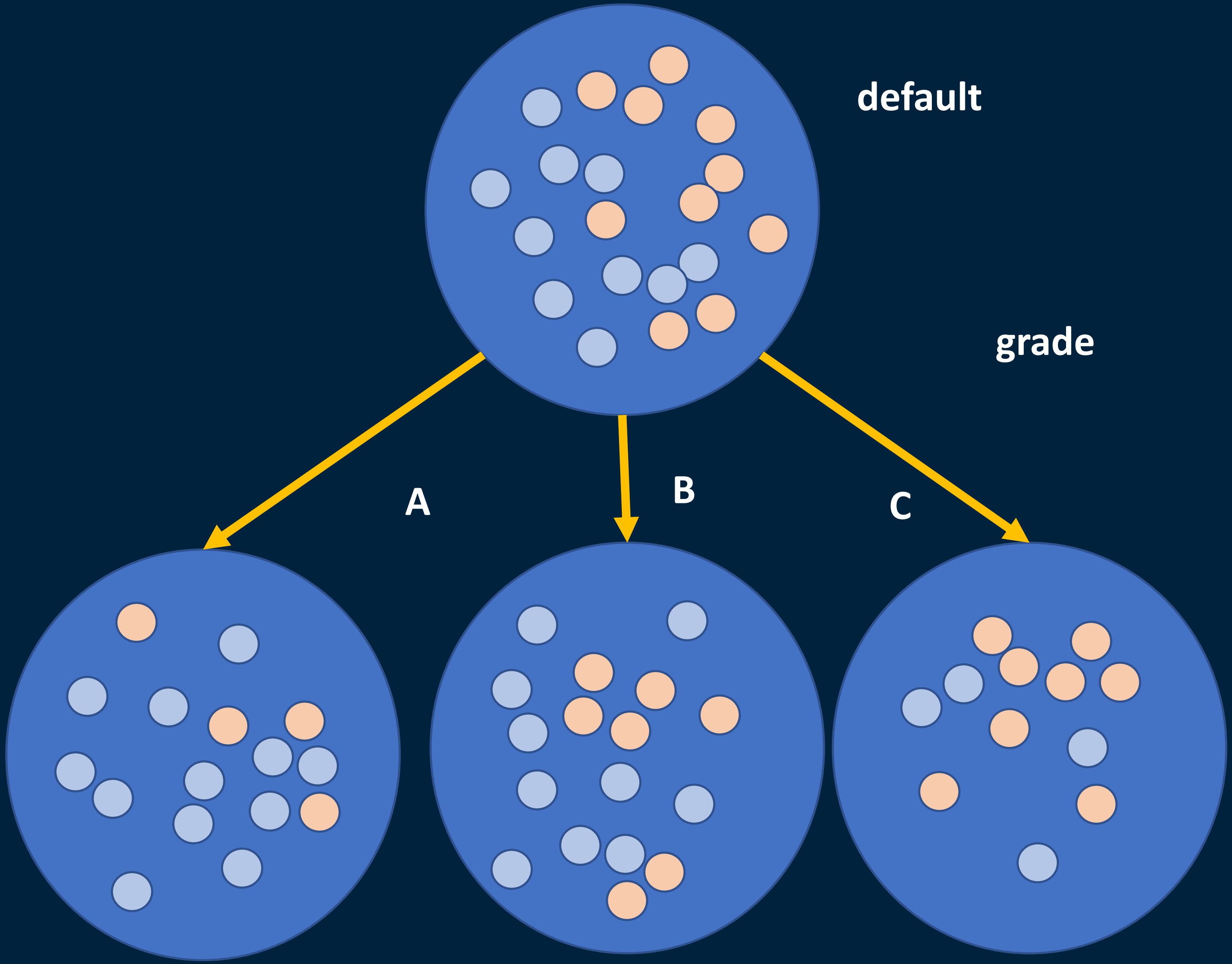
Common case

- The attributes separate imperfectly loans that will default and those that won’t
- In other words attributes will separate instances into “impure” sets
- There are many ways to separate into impure sets, how would we chose which one is best?



Entropy

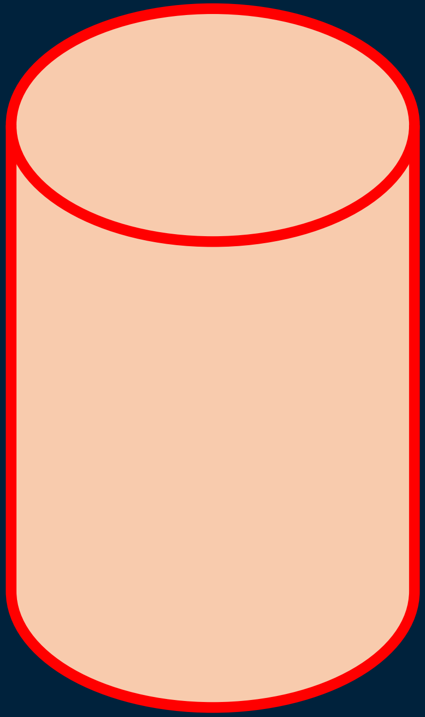
- Measures of impurity
- The notion of impurity is closely related to the notion of information



Entropy and Information

What is information?

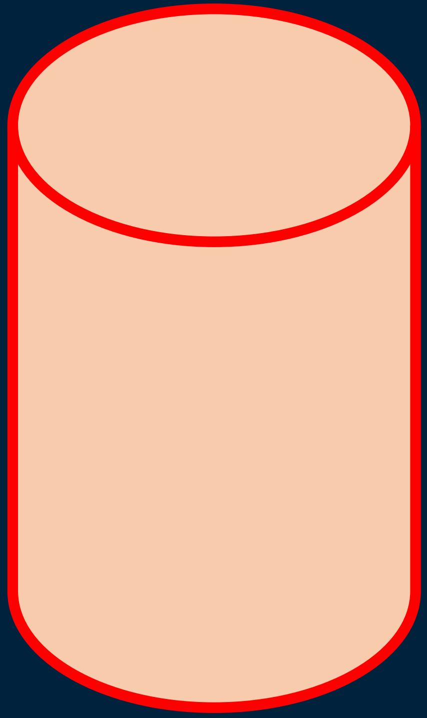
By definition, information is ***knowledge*** about things, which may or may not be conceived by an observer



Entropy and Information

What is information?

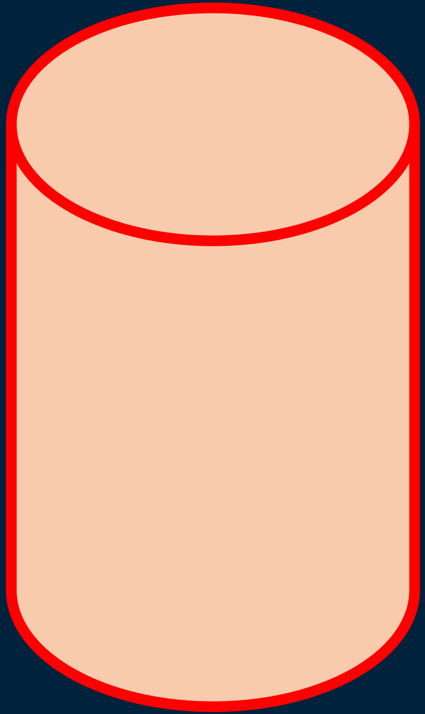
- How much information do we gain from learning the type of the ball drawn from the urn below?
- What did we learn that we didn't know before?



Entropy and Information

What is information?

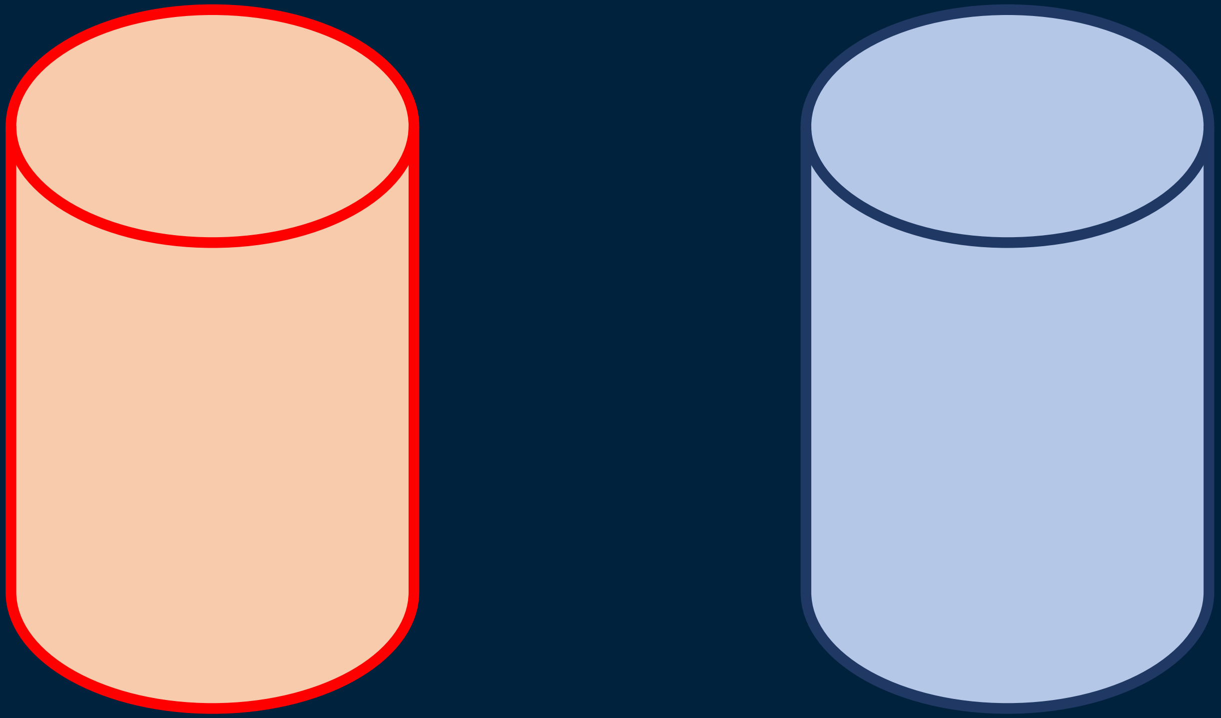
Information is the ***uncertainty*** of the outcome



Entropy and Information

What is information?

There is greater uncertainty on the type of the ball drawn from the red urn vs the blue urn, therefore we would say that a draw from the red urn has ***greater information content***



Information Measurement

Let H be a measure of information

- H should be ***maximized*** when the object is most unknown.
- ***H(X)=0*** if X is determined/certain
- The information measure H should be ***additive for independent objects***; i.e., with 2 information sources which has no relations with each other, H=H1+H2.

Entropy

- Entropy H(X) of a random variable X is defined by

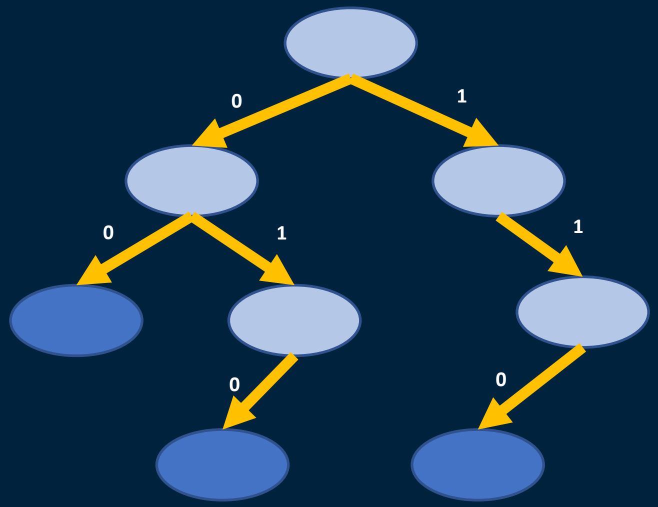
$$H(X)= -\sum p(x) \log p(x)$$

- We can verify that the measure H(X) satisfies the three criterion stated.
- If we choose the logarithm in base 2, then the entropy may be claimed to be in the unit of ***bits***; the use of the unit will be clarified later.

Information Measurement

Shannon Entropy

- How many bits on average would required to describe a path from the root to a leaf?



Entropy in thermodynamics or statistical mechanics

- Entropy is the measure of disorder of a thermodynamic system
- The definition is identical with the information entropy, but the summation now runs on all possible physical states
- Actually, entropy is first introduced in thermodynamics and Shannon found out his measure is just entropy in physics

Historical Notes

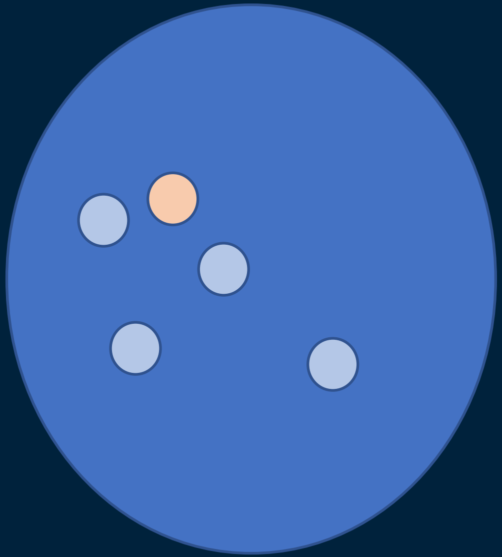
Claude E. Shannon (1916-2001) himself in 1948, has established almost everything we will talk about today.

He was dealing with communication aspects.

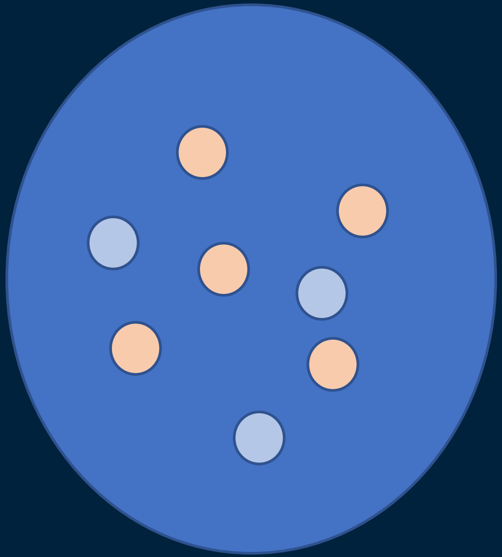
He first used the term “bit.”

Entropy

$$entropy = \sum_i p_i \cdot \log p_i$$

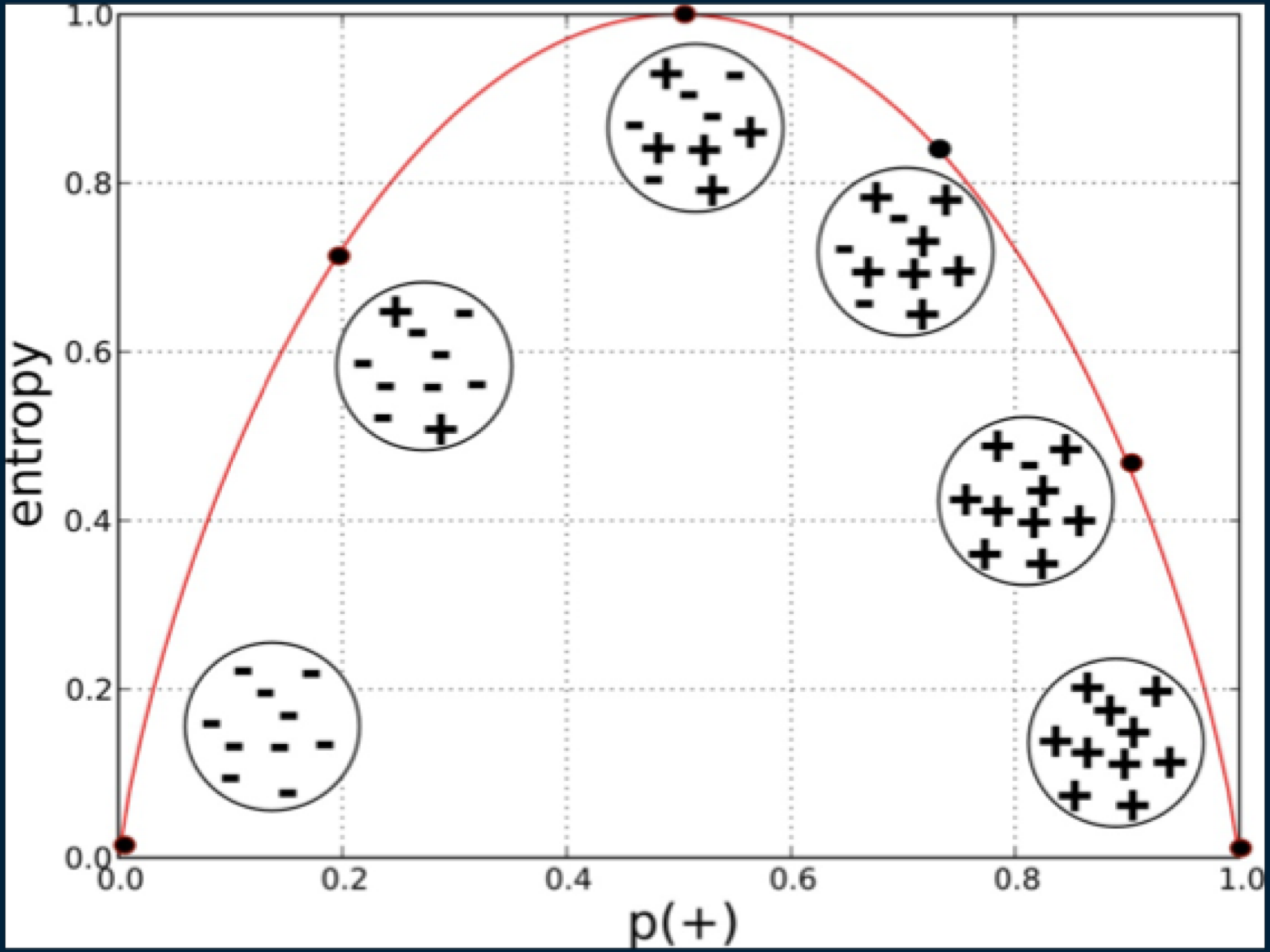


H=0.8*log 0.8+0.2*log 0.2=***0.7219***

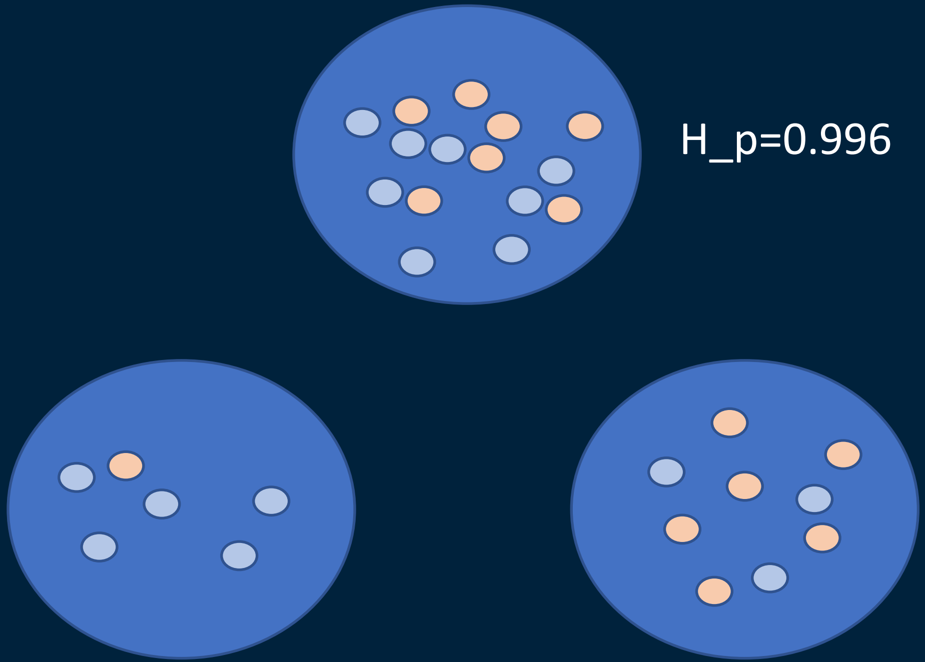


H=0.625*log 0.625+03752*log 0.375=***0.9544***

Entropy



Information Gain



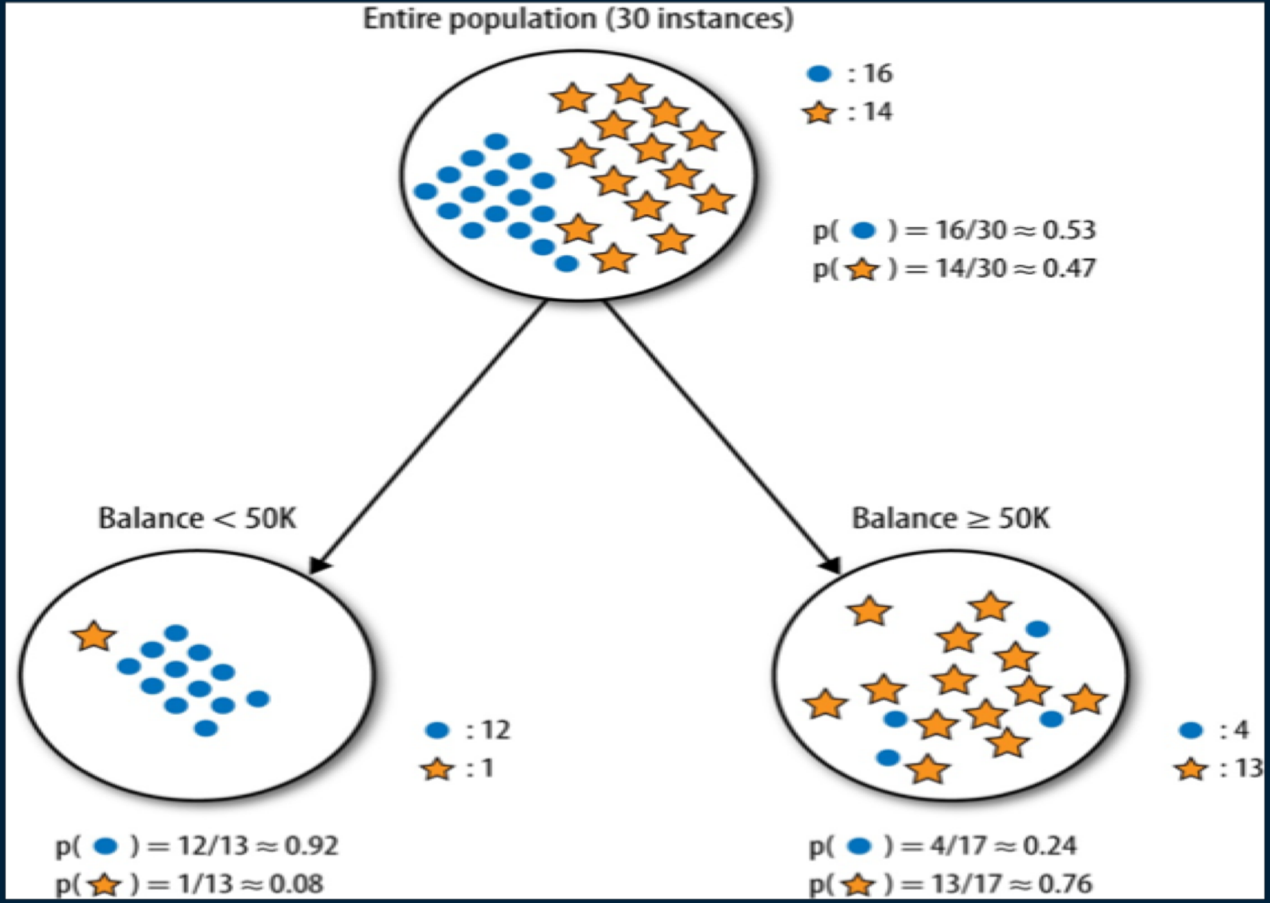
H_p=0.996

H_c=0.650

H_c=0.918

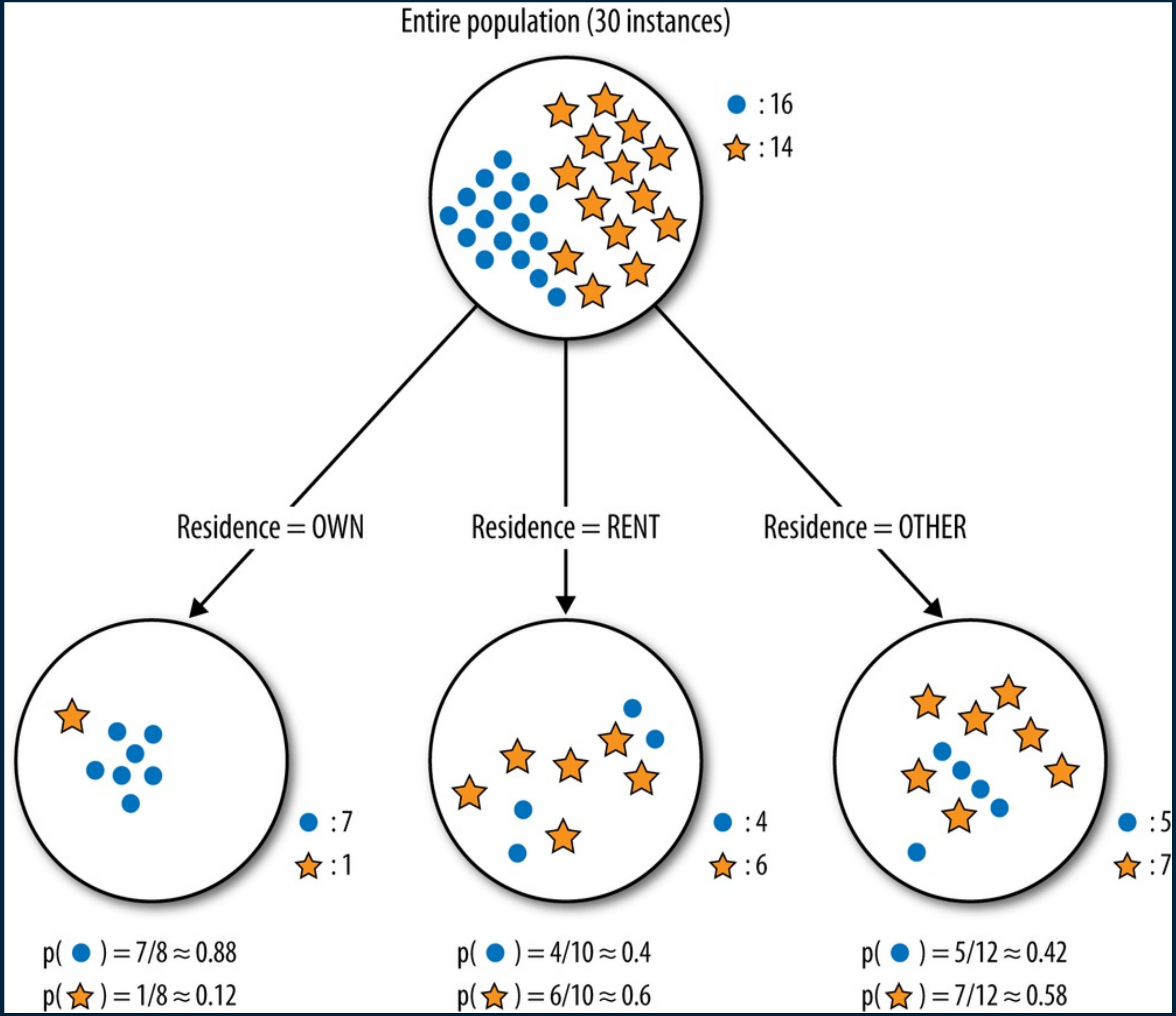
$$Ig = \sum p_{c_i} \cdot H_{c_1} - H_p = 0.1852$$

Information Gain



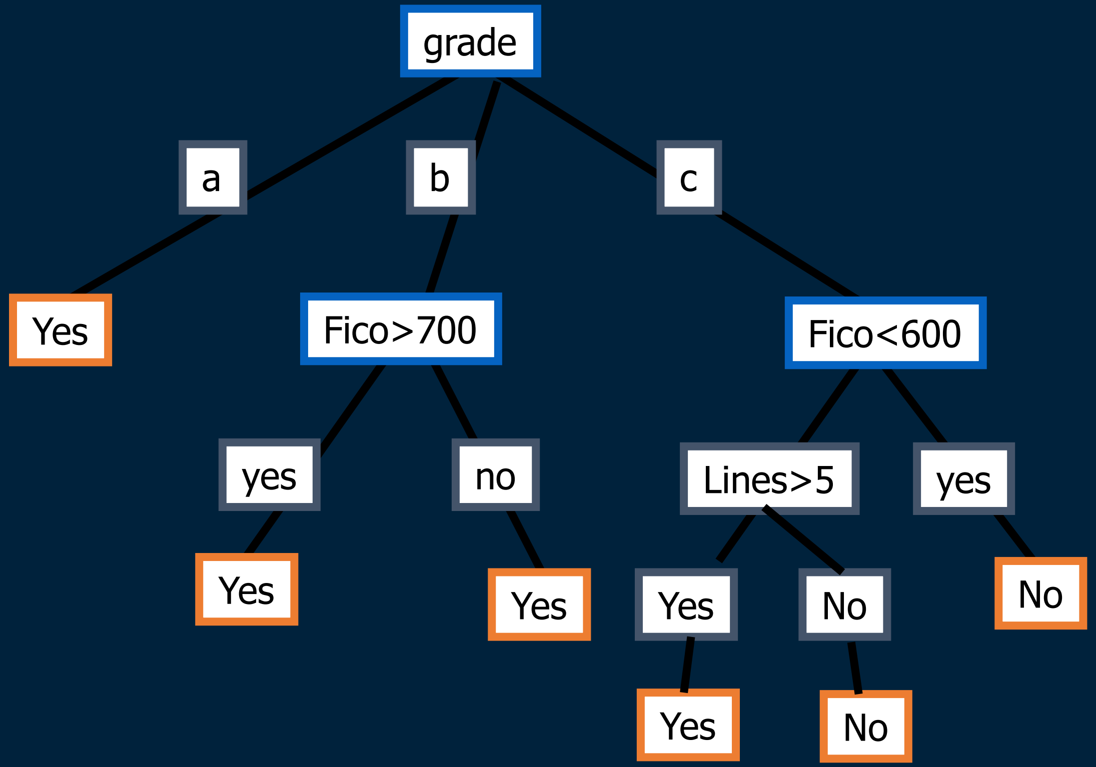
$$\begin{aligned}
entropy(parent) &\approx 0.99\\
entropy(\text{ Residence=OWN }) &\approx 0.54\\
entropy(\text{ Residence=RENT }) &\approx 0.97\\
entropy(\text{ Residence=OTHER }) &\approx 0.98\\
IG &\approx 0.13
\end{aligned}$$

Information Gain



ID3 algorithm

ID3 algorithm: Repeatedly find the split maximizing the information gain on the residual set



ID3 algorithm

- The decision tree represents the classification of the table
- It can classify all the objects in the table
- Each internal node represents a test on some property
- Each possible value of that property corresponds to a branch of the tree
- An individual of unknown type may be classified be traversing this tree

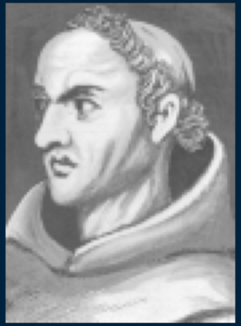
ID3 algorithm

- In classifying any given instance, the tree does not use all the properties in the table
- Decision tree for credit risk assessment
  - If a person has a good credit history and low debit, we ignore her collateral income and classify her as low risk
  - In spite of omitting certain tests, the tree classifies all examples in the table

ID3 algorithm

- ID3 algorithm assumes that a good decision tree is the simplest decision tree
- Heuristic:
  - Preferring simplicity and avoiding unnecessary assumptions
  - Known as Occam‘s Razor

Occam Razor



- Occam Razor was first articulated by the medieval logician William of Occam in 1324
  - born in the village of Ockham in Surrey (England) about 1285, believed that he died in a convent in Munich in 1349, a victim of the Black Death
  - It is vain do with more what can be done with less..
- We should always accept the simplest answer that correctly fits our data
- The smallest decision tree that correctly classifies all given examples